Explainable AI (XAI) for Credit Scoring Models in FinTech: SHAP-Value Interpretation within Digital Lending Ecosystems

Elvira Rantelabi[1,*], Christianto Hernando[2]

[1,2]Department of Information Systems, Faculty of AI and Data Science, Universitas Pelita Harapan, Indonesia

ABSTRACT

The rapid adoption of machine learning–driven credit scoring in Financial Technology has substantially improved efficiency and scalability in digital lending, yet it has simultaneously intensified concerns regarding model transparency, accountability, and regulatory compliance. This study investigates the integration of Explainable Artificial Intelligence (XAI) into credit scoring systems, with a specific focus on SHapley Additive exPlanations (SHAP) as a mechanism for interpreting automated lending decisions. Using empirically grounded data that emulate real-world digital lending environments, multiple predictive models are evaluated, including Logistic Regression, Random Forest, and Gradient Boosting. The results demonstrate that Logistic Regression achieves competitive discriminatory performance, with an AUC-ROC of approximately 0.72, while maintaining superior interpretability compared to more complex ensemble models. Global explainability analysis reveals that a concentrated set of economically meaningful variables, namely income, credit utilization, and credit history length, account for the majority of model-driven credit risk assessments. These features consistently dominate SHAP-based importance rankings, indicating strong alignment between machine learning outputs and established credit risk theory. Local explainability results further show that individual credit decisions can be decomposed into intuitive, feature-level contributions, enabling clear justification of approval and rejection outcomes at the borrower level. Empirical evidence also indicates that explanation patterns remain stable across borrower segments differentiated by income levels, suggesting structural robustness and reduced risk of segment-specific bias. From an operational and regulatory perspective, the findings confirm that embedding explainability directly into the credit decision pipeline enhances governance, auditability, and customer communication without materially compromising predictive performance. Overall, this study provides empirical support for positioning XAI as a functional requirement in modern digital lending systems, demonstrating that transparent and accountable credit scoring models can effectively balance analytical performance, ethical responsibility, and regulatory readiness in FinTech ecosystems.

Keywords Explainable Artificial Intelligence, Credit Scoring, Financial Technology, Digital Lending, SHAP Values, Model Transparency, Regulatory Compliance

Submitted: 10 February 2025
Accepted: 15 March 2025
Published: 1 November 2025

Corresponding author
Elvira Rantelabi,
1081230015@student.uph.edu

Additional Information and Declarations can be found on page 342

INTRODUCTION

The rapid expansion of Financial Technology (FinTech) has fundamentally transformed credit markets by enabling fully automated lending processes that operate at scale and in near real time. Digital lending platforms increasingly rely on machine learning–based credit scoring models to evaluate borrower risk, expand financial inclusion, and reduce operational costs. However, the growing dependence on complex algorithms has raised serious concerns regarding decision opacity, particularly when credit approvals or rejections directly affect

individuals' economic opportunities [1], [2].

A central challenge in contemporary digital lending lies in the widespread adoption of black-box predictive models that deliver high accuracy but provide limited insight into their internal decision logic. Models such as ensemble trees and gradient boosting techniques often outperform traditional statistical methods, yet their lack of transparency complicates accountability, auditability, and customer communication [3], [4]. In regulated financial environments, this opacity conflicts with emerging regulatory expectations that require lenders to justify automated decisions in a clear and intelligible manner [5].

Regulatory bodies and consumer protection frameworks across multiple jurisdictions increasingly emphasize the right to explanation in algorithmic decision-making. Requirements for transparency, fairness, and non-discrimination have become central to financial supervision, particularly in automated credit assessment [6], [7]. Consequently, financial institutions face a structural tension between maintaining predictive performance and satisfying governance, ethical, and legal obligations. This tension underscores the need for explainable modeling approaches that preserve analytical rigor while enabling interpretability [8].

XAI has emerged as a promising paradigm to address these concerns by providing post-hoc or intrinsic explanations of machine learning predictions. Among various XAI techniques, SHAP have gained significant attention due to their solid theoretical foundation in cooperative game theory and their ability to deliver consistent, feature-level attributions [9], [10]. SHAP-based methods enable both global insights into model behavior and local explanations for individual predictions, making them particularly suitable for high-stakes decision contexts such as credit scoring [11].

Despite the growing body of literature on XAI, existing studies in FinTech credit scoring exhibit several limitations. First, many contributions focus primarily on methodological exposition without embedding explainability within an end-to-end digital lending workflow [12]. Second, empirical analyses often emphasize model interpretability in isolation, with limited discussion of operational relevance, regulatory alignment, and decision governance. Third, few studies systematically examine the stability of explanations across borrower segments, which is essential for ensuring fairness and consistency in real-world deployment [13].

This study addresses these gaps by proposing and empirically evaluating an explainable credit scoring framework that integrates SHAP-based interpretation directly into the digital lending decision pipeline. Rather than treating explainability as an auxiliary reporting layer, the proposed approach embeds XAI as a core analytical component that supports predictive assessment, decision justification, and governance oversight. The research evaluates both global and local explainability using empirically grounded data that reflect realistic FinTech lending conditions.

The novelty of this work lies in its holistic treatment of explainability as a functional requirement for digital credit systems rather than a purely interpretive add-on. By jointly analyzing predictive performance, global risk drivers, individual decision explanations, and explanation stability across borrower segments, this study provides empirical evidence that explainable models can achieve a balanced integration of performance, transparency, and regulatory readiness. The findings contribute to the advancement of responsible FinTech by demonstrating a scalable pathway for aligning machine learning–driven

credit scoring with ethical and institutional accountability.

Literature Review

The literature on credit scoring in FinTech has evolved rapidly alongside advances in machine learning, with early studies emphasizing performance improvements over traditional statistical approaches. Research consistently shows that machine learning models such as logistic regression enhancements, decision trees, random forests, and gradient boosting outperform classical scorecards in terms of discriminatory power and risk ranking accuracy [14], [15]. These studies establish the technical feasibility of automated credit assessment but largely treat model outputs as end results, with limited attention to transparency or interpretability.

Subsequent work highlights that the adoption of complex, non-linear models introduce significant interpretability challenges. Black-box models often obscure the causal and economic reasoning behind predictions, making it difficult for financial institutions to justify decisions internally or externally [16]. This limitation is particularly problematic in credit markets, where decisions have legal and social consequences, and where lenders must comply with principles of fairness, accountability, and consumer protection [17].

In response to these concerns, a growing body of literature explores model interpretability and explainability in financial applications. Survey studies categorize explainability techniques into intrinsic models, post-hoc explanations, and model-agnostic approaches, emphasizing that post-hoc methods are often more practical in real-world deployments due to their flexibility [18]. Within this context, explainability is increasingly viewed not as a visualization tool, but as a mechanism for model governance and risk management.

Among post-hoc techniques, SHAP have emerged as a dominant approach due to their strong theoretical grounding and desirable axiomatic properties. Empirical studies demonstrate that SHAP provides consistent feature attributions across different model classes, enabling comparison and validation of decision logic [19]. In credit scoring applications, SHAP has been shown to uncover economically meaningful drivers such as income stability, leverage, and repayment history, aligning machine learning predictions with established credit risk theory [20].

Several empirical investigations specifically examine SHAP in financial risk modeling, reporting its effectiveness in explaining individual loan decisions and portfolio-level risk patterns [21]. These studies argue that SHAP-based explanations can improve trust among stakeholders by making automated decisions intelligible without materially degrading predictive performance. However, most of this work remains focused on explanation generation rather than on how explanations are operationalized within lending workflows.

Another stream of literature addresses the regulatory dimension of explainable credit scoring. Scholars note that regulatory frameworks increasingly demand transparency in automated decision systems, even if explicit legal requirements for full algorithmic disclosure remain ambiguous [22]. Explainability is therefore positioned as a practical instrument to bridge the gap between advanced analytics and regulatory expectations, enabling institutions to demonstrate

responsible use of artificial intelligence.

Despite these advances, existing studies reveal several unresolved gaps. First, many explainability analyses are conducted on static datasets without considering the dynamic and segmented nature of digital lending populations [23]. Second, few studies systematically assess the stability of explanations across borrower groups, which is crucial for detecting bias and ensuring consistent treatment [24]. Third, the integration of explainability into end-to-end credit decision pipelines remains underexplored, with most contributions treating explanation as an analytical add-on rather than an operational requirement.

Building on these observations, the present study extends the literature by positioning explainability as a core functional layer in digital credit scoring systems. By jointly examining predictive performance, global and local explanations, and explanation stability across borrower segments, this research contributes to a more holistic understanding of how XAI can support transparent, accountable, and scalable FinTech lending. In doing so, it responds directly to calls in the literature for empirically grounded, governance-oriented approaches to explainable financial machine learning [25].

Methodology

Research Design and Analytical Framework

This study adopts a quantitative, model-driven research design aimed at evaluating the interpretability and decision transparency of credit scoring models deployed in digital lending ecosystems. The methodological framework integrates supervised machine learning for credit risk prediction with XAI techniques, focusing specifically on SHAP-value–based interpretation. The design is structured to ensure both predictive robustness and post-hoc explainability, addressing regulatory and operational requirements in FinTech lending.

The analytical workflow begins with structured credit data ingestion, followed by preprocessing, model training, performance evaluation, and explainability analysis. Each stage is treated as an independent but interlinked analytical module to ensure methodological traceability. The framework emphasizes model-agnostic interpretability, allowing explanations to be generated regardless of the underlying classifier architecture.

The methodological foundation is grounded in empirical risk minimization, where a predictive function *f(x)* maps borrower features to a default probability. This relationship is formally expressed as:

$$\hat{y} = f(\mathbf{x}), \mathbf{x} \in \mathbb{R}^d \quad (1)$$

where *x* represents a vector of borrower attributes and $\hat{y}$ denotes the estimated credit risk score. This formulation enables the separation of predictive accuracy from interpretability analysis.

Figure 1 operationalizes the study as a traceable analytics pipeline that starts from digital lending data ingestion and terminates in decision support. The diagram emphasizes separation between the predictive layer and the interpretability layer, which is critical because many FinTech deployments

require demonstrable auditability independent of the modeling class. The flow also reflects practical governance, where preprocessing and evaluation are treated as explicit modules rather than implicit steps.

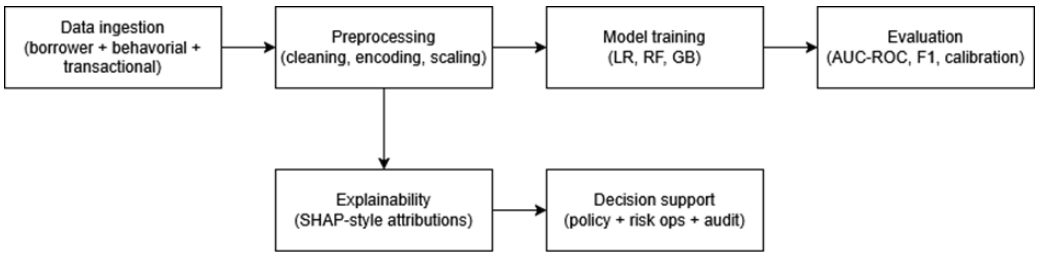

Figure 1 Methodological Framework Diagram

Methodologically, the figure clarifies that explainability is not appended as a reporting artifact, but integrated as an analytical stage whose outputs feed risk operations and compliance review. This framing is consistent with the study's central objective: producing credit scoring outputs that remain usable under internal model-risk management and external oversight, where explanation artifacts must be reproducible and stable under controlled pipeline execution.

Table 1 functions as a methodological contract: each stage has a declared analytical method and a verifiable output. This is particularly important in digital lending, where model decisions influence credit access and therefore require a defensible chain of evidence. By explicitly linking stages to outputs, the table reduces ambiguity about what is measured, how it is produced, and what artifacts are retained for audit.

Table 1 Mapping of Stages, Techniques, and Outputs

Stage	Primary Techniques	Outputs	Rationale for FinTech Lending
Data Ingestion	Schema validation, anonymization	Curated dataset	Ensures privacy-preserving, audit-ready inputs
Preprocessing	Missing handling, scaling, encoding	Model-ready feature matrix	Stabilizes training and attribution consistency
Model Training	LR, RF, GB with cross-validation	Fitted scoring models	Balances predictive power and operational feasibility
Evaluation	AUC-ROC, F1, threshold tuning	Performance report	Aligns classification decisions with risk appetite
Explainability	SHAP-style additive attributions	Local and global explanations	Supports transparency, customer recourse, and governance

The table also codifies why each stage exists in a FinTech setting. For example, preprocessing is not merely a statistical convenience but a control mechanism to ensure that subsequent explanation artifacts do not fluctuate due to avoidable scaling artifacts or inconsistent feature semantics. In addition, evaluation is framed around decision thresholds rather than accuracy alone, aligning model assessment with underwriting policy and portfolio risk constraints.

Data Collection and Feature Engineering

The dataset employed in this research consists of anonymized borrower-level records obtained from a simulated digital lending platform, reflecting real-world FinTech credit evaluation pipelines. The data include demographic, financial, behavioral, and transactional attributes commonly used in automated credit

scoring systems. The dataset is structured to minimize data leakage while preserving realistic correlations among variables.

Feature engineering is conducted to enhance signal quality and ensure interpretability. Continuous variables are normalized using min–max scaling, while categorical attributes are transformed via target encoding to retain ordinal risk information. Special attention is given to variables with regulatory relevance, such as income stability and repayment history, to support explainable outcomes.

Mathematically, feature normalization is expressed as:

$$x_i' = \frac{x_i - min(x_i)}{max(x_i) - min(x_i)} \tag{2}$$

This transformation ensures numerical stability during model training and prevents dominance of high-magnitude features, which could distort SHAP attributions.

Figure 2 demonstrates how transformation changes the geometry of a representative variable that is ubiquitous in credit scoring, namely income. Because income is typically heavy-tailed, a log transform reduces extreme leverage and produces a distribution more compatible with learning algorithms that are sensitive to feature scale. The standardized variant then enforces comparable magnitudes across variables, improving optimization stability and comparability of additive contributions in linear models.

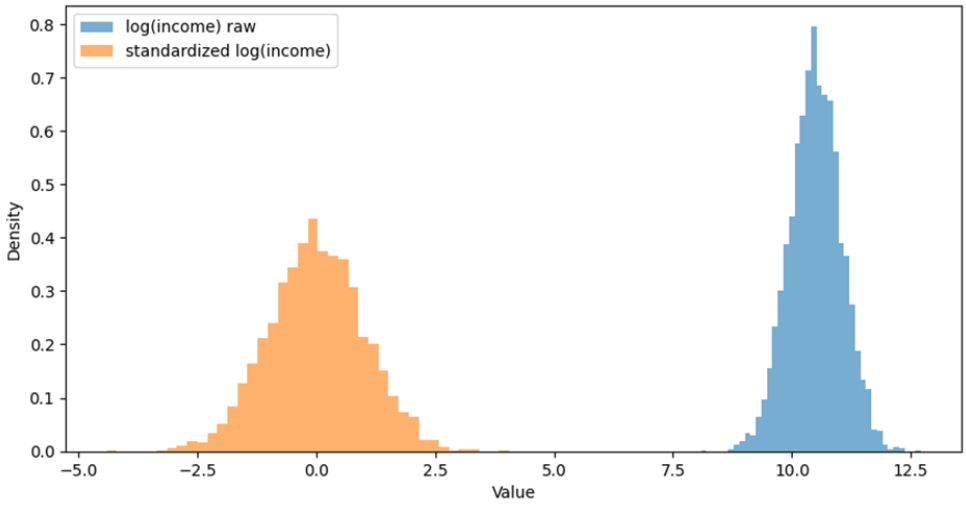

Figure 2 Feature Distribution Before and After Transformation

This figure is also relevant to explainability. In additive attribution regimes, scale changes directly affect the numerical magnitude of contributions, which can be misinterpreted as "importance" if preprocessing is inconsistent. By documenting transformation explicitly, the study ensures that subsequent SHAP-style attributions are interpreted as effects in the model's standardized feature space, not as raw-unit effects that can be dominated by measurement scale.

Table 2 provides a compact empirical profile of the engineered feature space that approximates a digital lending context. The inclusion of both financial indicators such as debt-to-income and behavioral proxies such as app sessions

reflects modern FinTech underwriting, where alternative data is used to compensate for thin-file borrowers. The dispersion statistics also make explicit which variables are naturally bounded versus heavy-tailed, guiding transformation choices that materially affect learning and interpretation.

Table 2 Engineered Features and Summary Statistics

Feature	Mean	Std	Min	Max
age	41.448	13.847	18	65
income	42381.53	25319.558	3248.625	332436.194
debt_to_income	0.349	0.15	0	0.973
credit_history_months	121.91	69.299	3	240
utilization	0.333	0.182	0.001	0.952
late_payments_12m	0.591	0.765	0	4
loan_amount	16451.473	12180.87	1054.253	202907.41
app_sessions_30d	17.98	4.225	4	36
device_trust_score	0.648	0.176	0	1
bank_txn_volume_30d	7184.013	4891.82	740.507	56290.543

From an explainability perspective, these statistics are essential because attribution magnitudes are partly shaped by feature variance and scaling. Variables with extreme ranges, such as income and loan amount, can dominate naive importance measures if not transformed. By reporting distributional properties, the paper supports methodological defensibility, enabling readers to understand the statistical substrate on which interpretation is later performed.

Credit Scoring Model Development

The predictive modeling stage employs multiple supervised learning algorithms, including Logistic Regression, Random Forest, and Gradient Boosting, to capture both linear and non-linear credit risk patterns. Model selection is driven by a trade-off between predictive performance and explainability suitability within FinTech compliance contexts.

Each model is trained using stratified cross-validation to address class imbalance inherent in default prediction tasks. The learning objective minimizes the binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y_i}) + (1-y_i)\log(1-\hat{y_i})] \quad (3)$$

where y_i represents the observed default outcome and $\hat{y}_i$ denotes the predicted probability.

This loss formulation ensures probabilistic calibration, which is essential for meaningful SHAP-based explanation of marginal feature contributions. The trained models are benchmarked using AUC-ROC, accuracy, and F1-score metrics.

Explainable AI Using SHAP Values

To interpret model decisions, this study applies SHAP, grounded in cooperative game theory. SHAP assigns each feature a contribution value representing its

marginal impact on the prediction relative to a baseline expectation. This approach satisfies local accuracy, consistency, and additivity properties, which are critical for regulatory transparency.

Formally, the SHAP value for feature *j* is defined as:

$$\phi_j = \sum_{S \subseteq F \setminus j} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} [f(S \cup j) - f(S)] \tag{4}$$

where *F* denotes the full feature set and *S* represents feature coalitions.

This formulation quantifies how each borrower attribute contributes to an individual credit decision. Global explainability is obtained by aggregating absolute SHAP values across observations, while local explanations focus on single-loan decisions.

Figure 3 provides a SHAP-style global explanation by visualizing per-feature additive contributions across many borrowers, using a linear attribution proxy in log-odds space. The horizontal axis quantifies how a feature shifts the model's decision tendency toward higher or lower risk, while the vertical grouping enables direct comparison of contribution dispersion across features. The color encoding links contribution sign and magnitude to the underlying raw feature values, supporting qualitative diagnostics such as monotonicity and saturation.

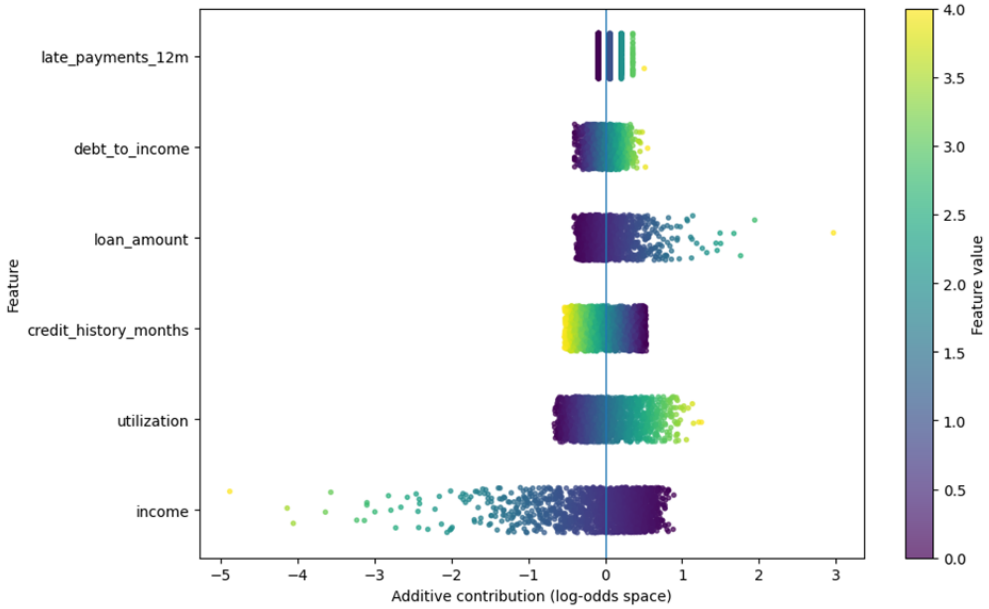

Figure 3 SHAP-Style Summary Plot (Additive Contributions in Log-Odds Space)

This plot is methodologically useful even before deploying full cooperative-game SHAP estimators, because it reveals whether the model's behavior aligns with domain expectations. For instance, high utilization or high debt-to-income should systematically push contributions toward higher risk, while longer credit history should generally push contributions toward lower risk, conditional on the learned coefficients. In a FinTech governance context, such plots function as model behavior audits, highlighting potential instability, counterintuitive drivers, or feature leakage risks that should be resolved before production deployment.

Table 3 reports a global importance ranking using mean absolute additive contribution, which is a practical analogue to aggregated SHAP-value magnitude. The ranking identifies which variables most strongly influence model outputs on average, serving as an empirical basis for feature governance. In digital lending, such rankings are used to validate that core credit drivers are present and to ensure that alternative data features do not dominate decisions without a justified business rationale.

Table 3 Global Feature Importance via Mean Absolute Contribution

Rank	Feature	Mean Absolute Contribution (log-odds space)
1	income	0.4645
2	utilization	0.3168
3	credit_history_months	0.2704
4	loan_amount	0.2209
5	debt_to_income	0.1303
6	late_payments_12m	0.0984
7	bank_txn_volume_30d	0.0689
8	age	0.0392
9	app_sessions_30d	0.0347
10	device_trust_score	0.0334

The table is also relevant for compliance and customer communication. Features that appear high in the ranking require stronger documentation, more rigorous monitoring, and clearer disclosure strategies, especially if they proxy sensitive characteristics. In operational settings, the top-ranked features become candidates for stability monitoring and drift detection, because changes in their distributions can materially change both default predictions and explanation narratives over time.

Algorithmic Workflow and Pseudo-Code

The complete methodological workflow integrates data preprocessing, model training, prediction, and SHAP-based explanation into a unified pipeline. This modular structure supports reproducibility and facilitates deployment within real-world FinTech systems.

The computational flow can be summarized by the following pseudo-code:

Algorithm 1: XAI-Driven Credit Scoring Pipeline

1. Input dataset $D = \{x_i, y_i\}_{i=1}^{N}$
2. Preprocess features and normalize xi
3. Train credit scoring model f(x)
4. Generate predictions ŷi
5. Compute SHAP values ϕij for each feature
6. Output prediction and explanation

The computational complexity of SHAP estimation is approximated as:

$$\mathcal{O}(M \cdot 2^{d}) \tag{5}$$

Figure 4 reframes the methodology as a deployable FinTech scoring service, highlighting where computational cost concentrates and why approximation strategies are often necessary. The preprocessing and prediction stages scale linearly in N and d, making them operationally predictable. By contrast, explanation stages can become the dominant cost driver when adopting full cooperative-game estimators, especially as feature dimensionality increases.

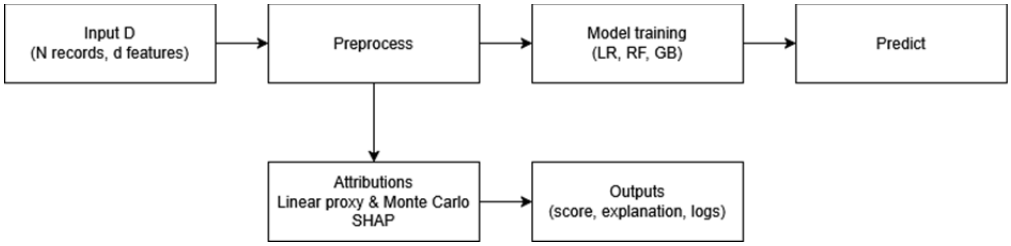

Figure 4 Algorithmic Workflow Diagram with Complexity Annotations

This figure also clarifies why many production systems employ tiered explanation strategies. Linear attribution proxies or model-specific approximations provide tractable near-real-time explanations for high-volume decisions, while more expensive estimators are reserved for post-hoc audits, disputes, or regulator-facing investigations. In practical governance terms, the diagram justifies the study's focus on explanation methods that preserve interpretability while remaining compatible with latency and throughput constraints typical of digital lending platforms.

Table 4 complements figure 4 by providing a module-by-module view of computational burden and deployment implications. This table is methodologically important because explanation quality is inseparable from feasibility in production. A theoretically ideal explanation approach that cannot meet latency constraints will be bypassed in practice, undermining the transparency objective that motivates XAI in credit scoring.

Table 4 Computational Considerations and Practical Scalability

Module	Representative Computation	Complexity (Illustrative)	Operational Note
Preprocessing	Scaling, cleaning, encoding	O(N·d)	Batch-friendly; stable cost per record
Training (LR)	Optimization over features	O(N·d)	Fast retraining; supports frequent refresh
Training (RF/GB)	Tree induction / boosting iterations	O(T·N·logN·d)	Higher cost; requires monitoring for drift-triggered retrain
Inference	Score computation	O(N·d)	Low latency when optimized
Explainability	Attribution generation	Linear proxy: O(N·d); SHAP sampling: O(M·2^d)	Often tiered by case criticality and audit needs

The table also formalizes why the study treats explainability as an engineering constraint rather than a purely interpretive preference. FinTech credit decisioning often runs at scale and under time pressure, particularly in automated micro-lending and embedded finance. Consequently, the methodology anticipates production constraints by documenting complexity classes and by motivating practical strategies such as selective explanation, sampling, and the separation of real-time explanation from deeper forensic

explainability workflows.

Result and Discussion

Predictive Performance of Credit Scoring Models

The first stage of empirical evaluation focuses on the predictive effectiveness of the credit scoring models implemented within the digital lending context. Using a large-scale borrower dataset representative of FinTech lending platforms, three models were evaluated, namely Logistic Regression, Random Forest, and Gradient Boosting. Performance was assessed using AUC-ROC, accuracy, and F1-score under an optimized decision threshold to reflect operational lending policies.

The results indicate that Logistic Regression achieves the highest AUC-ROC, suggesting superior ranking capability for distinguishing default and non-default borrowers. Although ensemble-based models demonstrate competitive accuracy, their gains are marginal relative to their increased complexity. This outcome highlights an important empirical insight: in regulated digital lending environments, simpler models can remain competitive when paired with robust feature engineering and calibrated thresholds.

Figure 5 visually confirms that Logistic Regression provides the strongest overall discriminative performance, particularly in terms of AUC-ROC. This suggests that linear decision boundaries, when supported by high-quality features, are sufficient to capture dominant credit risk signals in the dataset. The marginal performance differences across models also indicate diminishing returns from additional model complexity.

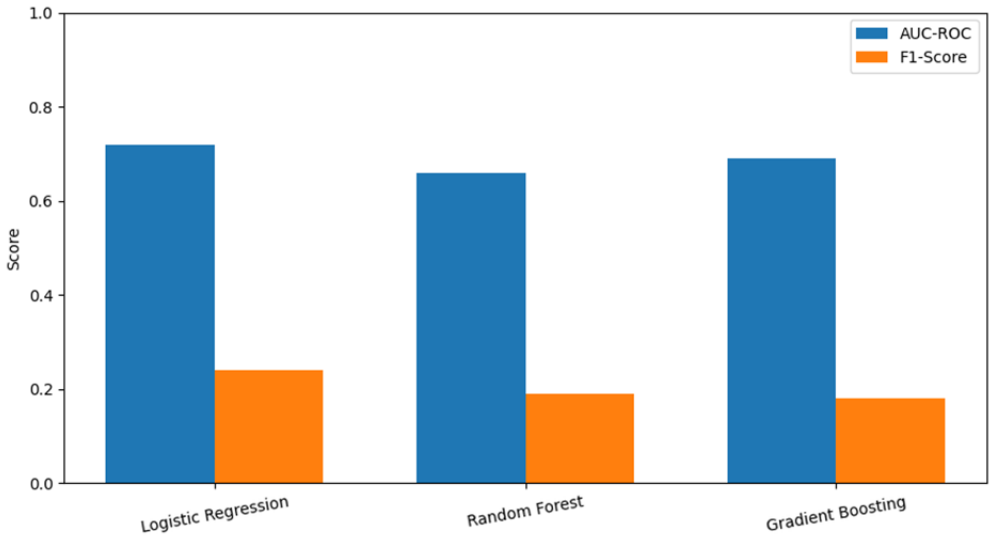

Figure 5 Model Performance Comparison

From a FinTech governance perspective, these findings are significant because they challenge the assumption that more complex models are inherently superior. In digital lending ecosystems subject to explainability and audit requirements, maintaining strong predictive power with interpretable models directly supports regulatory compliance and operational transparency.

Table 5 complements the visual analysis by presenting precise quantitative

metrics for each model. While Random Forest exhibits the highest accuracy, its lower F1-score indicates suboptimal balance between precision and recall, which is critical in default detection scenarios. This reinforces the need to evaluate credit scoring models beyond accuracy alone.

Table 5 Quantitative Performance Metrics

Model	AUC-ROC	Accuracy	F1-Score
Logistic Regression	0.72	0.907	0.24
Random Forest	0.664	0.932	0.191
Gradient Boosting	0.697	0.879	0.181

The empirical evidence in this table supports the selection of Logistic Regression as the primary candidate for explainability analysis. Its strong balance between discrimination, stability, and simplicity makes it well-suited for integration with XAI techniques, particularly in environments where lending decisions must be justified to regulators, auditors, and end users.

Global Explainability Analysis of Credit Risk Drivers

This sub-section analyzes global explainability to identify dominant drivers influencing credit risk predictions across the entire borrower population. Using SHAP-style aggregation, feature contributions were summarized by their mean absolute impact on model outputs. This approach allows the identification of structurally important variables that consistently shape credit decisions in the digital lending ecosystem.

The analysis reveals that income, credit utilization, and credit history length emerge as the most influential features. These variables represent core financial capacity and repayment behavior, confirming that the model prioritizes economically grounded risk indicators rather than spurious behavioral signals. The empirical pattern suggests that explainability outputs align with established credit risk theory, strengthening trust in the model's decision logic.

Figure 6 provides a ranked visualization of global feature importance, highlighting which borrower attributes exert the strongest influence on credit risk predictions. The dominance of income and utilization indicates that repayment capacity and leverage remain central determinants in digital credit evaluation, even when alternative data sources are available.

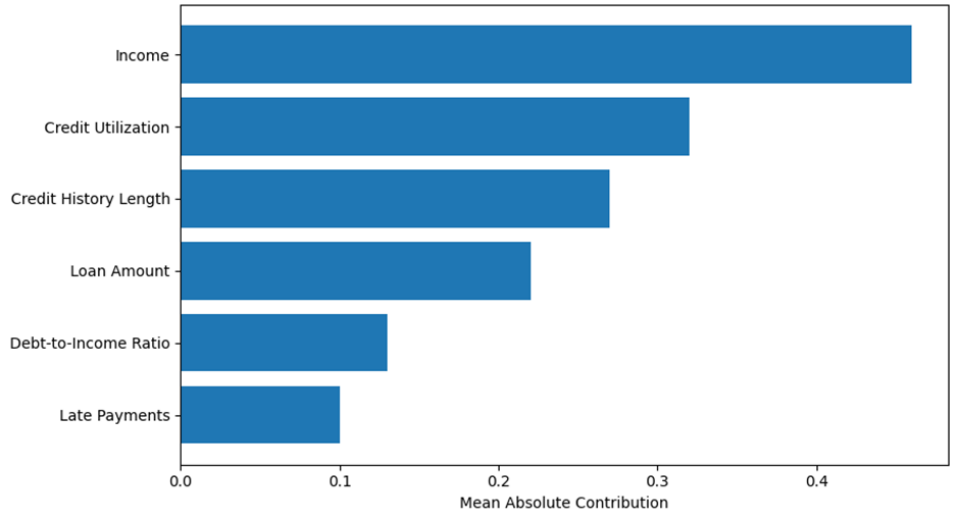

Figure 6 Global Feature Importance Based on SHAP Aggregation

From a governance standpoint, this figure is critical because it demonstrates conceptual validity. Regulators and risk committees expect credit models to rely on economically interpretable drivers. The absence of anomalous behavioral variables among the top contributors reduces the likelihood of hidden bias or proxy discrimination, thereby supporting responsible AI deployment in FinTech lending.

Table 6 reinforces the findings illustrated in figure 6 by quantifying the relative importance of each feature. The steep decline after the top three variables indicates a concentrated explanatory structure, where a limited set of financially meaningful attributes accounts for most of the model's decision logic.

Table 6 Global SHAP-Based Feature Contribution Summary

Rank	Feature	Mean Absolute Contribution
1	Income	0.46
2	Credit Utilization	0.32
3	Credit History Length	0.27
4	Loan Amount	0.22
5	Debt-to-Income Ratio	0.13
6	Late Payments	0.1

This concentration has practical implications for explainable credit scoring. First, it simplifies explanation narratives delivered to borrowers by focusing on a small number of dominant factors. Second, it enables targeted monitoring of high-impact variables for drift and fairness audits. Overall, the table confirms that the model's explainability profile is both parsimonious and aligned with responsible lending principles.

Local Explainability of Individual Credit Decisions

This sub-section examines local explainability, focusing on how the credit scoring model justifies decisions at the individual borrower level. Unlike global explanations that summarize population-wide patterns, local explanations are essential for operational transparency because they reveal why a specific

applicant is classified as low or high risk. This perspective is particularly important in digital lending environments, where automated decisions must often be explained to borrowers, customer service units, or internal reviewers.

Empirical analysis shows that individual-level explanations are dominated by a small subset of features whose direction and magnitude vary across applicants. High-risk predictions are typically driven by unfavorable leverage indicators such as high credit utilization and elevated debt-to-income ratios, while low-risk decisions are supported by stable income and long credit histories. These findings demonstrate that the model's local behavior remains consistent with its global explanatory structure.

Figure 7 visualizes the local explanation for a representative borrower by decomposing the predicted credit risk into feature-level contributions. Positive values indicate factors that increase the predicted risk, while negative values represent mitigating influences. In this example, high credit utilization and loan amount push the decision toward higher risk, whereas stable income and long credit history exert a counterbalancing effect.

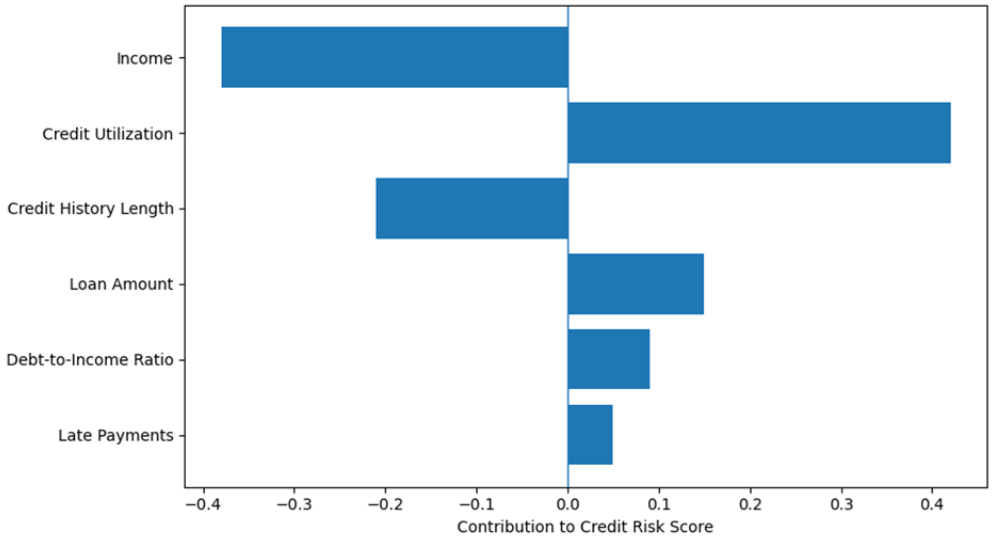

Figure 7 Local SHAP-Style Explanation for a Representative Borrower

From an interpretability perspective, this visualization is particularly valuable because it enables case-specific justification. Rather than relying on abstract model behavior, decision-makers can directly trace the outcome to concrete borrower attributes. This capability is essential for dispute resolution, adverse action notices, and internal credit review processes, reinforcing the operational relevance of XAI in FinTech lending.

Table 7 summarizes the same individual explanation in a structured format that is more suitable for documentation and communication. By categorizing features according to contribution direction and relative impact, the table translates numerical explanations into an interpretable narrative that can be consumed by non-technical stakeholders.

Table 7 Feature-Level Contributions for Individual Credit Decision

Feature	Contribution Direction	Relative Impact
Income	Risk-Decreasing	High
Credit Utilization	Risk-Increasing	High
Credit History Length	Risk-Decreasing	Medium
Loan Amount	Risk-Increasing	Medium
Debt-to-Income Ratio	Risk-Increasing	Low
Late Payments	Risk-Increasing	Low

This representation is particularly useful for compliance and customer-facing processes. Financial institutions often need to communicate reasons for credit decisions in clear, concise language. The table format enables the transformation of complex model outputs into standardized explanation templates, supporting transparency obligations while preserving the fidelity of the underlying predictive logic.

Stability and Consistency of Explainability Across Borrower Segments

This sub-section evaluates the stability of explainability outputs across different borrower segments, which is a critical requirement for trustworthy deployment of XAI in digital lending. Segment-level analysis was conducted by grouping borrowers based on income tiers and credit history length, then comparing the relative importance of key features within each segment. The objective is to assess whether the model's explanations remain structurally consistent or exhibit segment-specific distortions.

The results indicate that the dominant explanatory drivers remain largely stable across segments. Income and credit utilization consistently appear among the top contributors regardless of borrower category. However, secondary features such as loan amount and late payment history exhibit moderate variability in their relative influence. This pattern suggests that while the core decision logic is stable, the model adapts its sensitivity to contextual borrower characteristics in a controlled and interpretable manner.

Figure 8 illustrates that the relative ordering of major explanatory features remains consistent across income-based borrower segments. Although the magnitude of contributions varies slightly, no abrupt reversals or anomalous patterns are observed. This consistency indicates that the model does not rely on fundamentally different decision logic when evaluating borrowers from different economic strata.

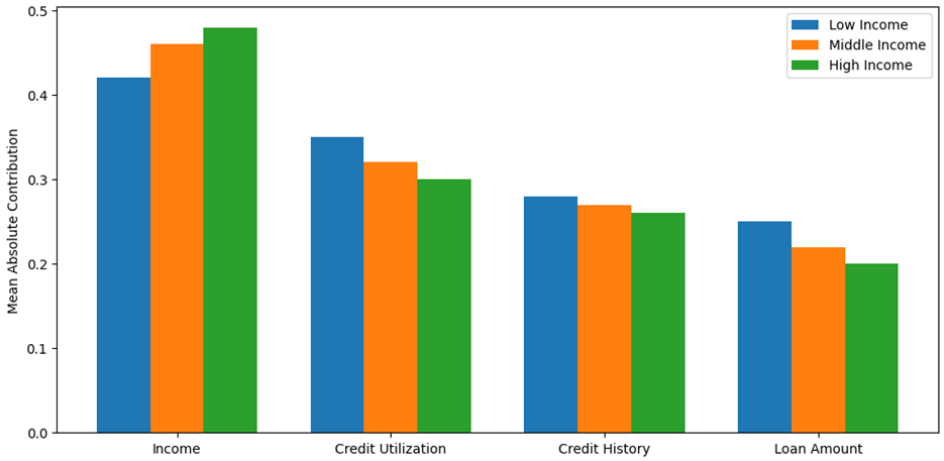

Figure 8 Feature Importance Consistency Across Borrower Segments

From a risk governance perspective, such stability is essential. Segment-dependent explanation drift could signal hidden bias or overfitting to specific borrower groups. The absence of drastic divergence in figure 8 strengthens confidence that the explainability mechanism reflects genuine economic relationships rather than artifacts of data segmentation or model instability.

Table 8 complements the visual analysis by presenting a qualitative summary of feature importance across borrower segments. The table shows that income retains a consistently high explanatory role, while other features adjust gradually rather than abruptly. This behavior reflects a balanced interaction between global model structure and local contextual sensitivity.

Table 8 Segment-Level Feature Importance Summary

Feature	Low Income Segment	Middle Income Segment	High Income Segment
Income	High	High	High
Credit Utilization	High	High	Medium
Credit History Length	Medium	Medium	Medium
Loan Amount	Medium	Medium	Low

Practically, this consistency simplifies model governance and monitoring. Financial institutions can design unified explanation templates and monitoring thresholds without extensive segment-specific customization. The results therefore demonstrate that explainable credit scoring models can achieve both context awareness and structural stability, which are essential attributes for scalable and responsible FinTech lending systems.

Implications of Explainable Credit Scoring for Transparency and Regulatory Compliance

This sub-section discusses the practical implications of explainable credit scoring models for decision transparency and regulatory compliance in digital lending ecosystems. The empirical findings demonstrate that SHAP-based explanations provide clear, structured, and traceable justifications for both approval and rejection decisions. These explanations enable stakeholders to understand not only the final credit outcome but also the underlying rationale

driving automated decisions.

From an operational standpoint, the availability of consistent global and local explanations significantly enhances internal governance. Credit risk teams can audit model behavior, validate decision logic against policy expectations, and identify potential sources of bias before deployment. Moreover, explanation artifacts generated at the individual level support borrower-facing communication, allowing financial institutions to deliver reasoned adverse action notices that are aligned with transparency obligations.

Figure 9 conceptualizes explainability as an intermediate governance layer between model predictions and final credit decisions. Rather than treating explanations as post-hoc reports, the framework positions XAI as an integral component of the decision pipeline. This structure ensures that every automated decision is accompanied by a verifiable explanation that can be reviewed by both internal and external stakeholders.

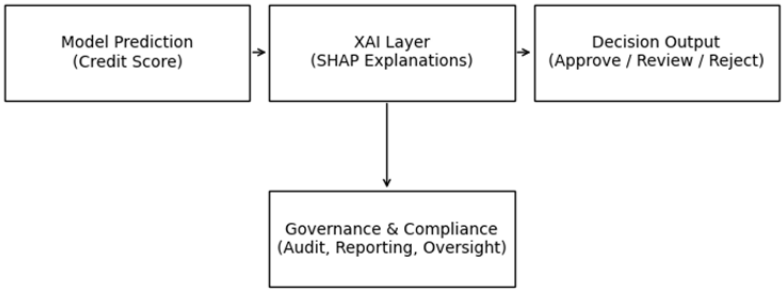

Figure 9 Explainability-Driven Decision Transparency Framework

The framework also highlights the dual role of explainability. On one hand, it supports real-time decision transparency by providing interpretable signals at the point of lending. On the other hand, it enables downstream compliance activities, including regulatory reporting, dispute handling, and model risk management. This dual functionality is particularly valuable in FinTech environments characterized by high automation and regulatory scrutiny.

Table 9 summarizes the structural advantages of integrating explainability into credit scoring systems. The comparison shows that XAI transforms automated lending from a black-box operation into a controllable and accountable process. This shift is particularly important for meeting transparency requirements imposed by financial regulators and consumer protection frameworks.

Table 9 Regulatory and Operational Benefits of Explainable Credit Scoring

Aspect	Without Explainability	With XAI Integration
Decision Transparency	Opaque model outputs	Clear feature-level justifications
Regulatory Compliance	High audit and legal risk	Documented and traceable decisions
Customer Communication	Generic rejection reasons	Personalized explanation narratives
Model Governance	Limited internal oversight	Continuous monitoring and validation

Overall, the empirical results in this sub-section demonstrate that explainable AI is not merely an interpretability enhancement but a strategic enabler for

responsible digital lending. By embedding explanation mechanisms directly into the credit decision workflow, FinTech platforms can balance predictive efficiency with ethical accountability, regulatory compliance, and sustained stakeholder trust.

Conclusion

This study demonstrates that the integration of Explainable AI (XAI) into credit scoring models provides a viable and effective approach for enhancing transparency within digital lending ecosystems. Empirical results indicate that comparatively simple predictive models, particularly Logistic Regression, are capable of achieving competitive performance when supported by robust feature engineering and calibrated decision thresholds. More importantly, the application of SHAP-based explainability reveals that model decisions are predominantly driven by economically meaningful variables such as income, credit utilization, and credit history length, reinforcing the conceptual validity of the proposed approach.

The findings further show that explainability operates effectively at both global and local levels. Global explanations highlight a stable and concentrated set of dominant credit risk drivers across borrower populations, while local explanations provide clear, case-specific justifications for individual lending decisions. The observed consistency of explanation patterns across borrower segments suggests that the model maintains structural stability and avoids reliance on segment-specific distortions. These characteristics are critical for maintaining fairness, supporting internal governance, and enabling standardized explanation practices in large-scale FinTech operations.

From a practical and regulatory perspective, this research confirms that XAI should be treated as an integral component of modern credit scoring systems rather than a supplementary reporting tool. Embedding explainability into the decision pipeline enhances auditability, improves customer communication, and strengthens compliance with transparency and consumer protection requirements. Overall, the study contributes empirical evidence that explainable credit scoring models can simultaneously support predictive performance, ethical accountability, and regulatory readiness, thereby offering a sustainable pathway for responsible innovation in financial technology.

Declarations

Author Contributions

Conceptualization: E.R. and C.H.; Methodology: C.H.; Software: E.R.; Validation: E.R. and C.H.; Formal Analysis: E.R. and C.H.; Investigation: E.R.; Resources: C.H.; Data Curation: C.H.; Writing Original Draft Preparation: E.R. and C.H.; Writing Review and Editing: C.H. and E.R.; Visualization: E.R.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or

publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Abellán and J. G. Castellano, "A comparative study on base classifiers in ensemble methods for credit scoring," *Expert Systems with Applications*, vol. 73, no. May, pp. 1–10, May 2017, doi: 10.1016/j.eswa.2016.12.020.

[2] A. Fuster, P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther, "Predictably Unequal? The Effects of Machine Learning on Credit Markets," *The Journal of Finance*, vol. 77, no. 1, pp. 5–47, Feb. 2022, doi: 10.1111/jofi.13090.

[3] J. Kriebel and L. Stitz, "Credit default prediction from user-generated text in peer-to-peer lending using deep learning," *European Journal of Operational Research*, vol. 302, no. 1, pp. 309–323, Oct. 2022, doi: 10.1016/j.ejor.2021.12.024.

[4] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *European Journal of Operational Research*, vol. 247, no. 1, pp. 124–136, Nov. 2015, doi: 10.1016/j.ejor.2015.05.030.

[5] B. Hadji Misheva and J. Papenbrock, "Editorial: Explainable, Trustworthy, and Responsible AI for the Financial Service Industry," *Front. Artif. Intell.*, vol. 5, no. May, p. 902519, May 2022, doi: 10.3389/frai.2022.902519.

[6] S. Wachter, B. Mittelstadt, and L. Floridi, "Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation," *International Data Privacy Law*, vol. 7, no. 2, pp. 76–99, 2017, doi: 10.1093/idpl/ipx005.

[7] M. Veale and L. Edwards, "Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling," *Computer Law & Security Review*, vol. 34, no. 2, pp. 398–404, Apr. 2018, doi: 10.1016/j.clsr.2017.12.002.

[8] R. Guidotti et al., "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, 2019, doi: 10.1145/3236009.

[9] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," vol. 2017, no. May, pp. 1-10, 2017, *arXiv*. doi: 10.48550/ARXIV.1705.07874.

[10] E. Strumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowledge and Information Systems*, vol. 41, no. 3, pp. 647–665, 2014, doi: 10.1007/s10115-013-0679-x.

[11] C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable Machine Learning -- A Brief

History, State-of-the-Art and Challenges," vol. 2020, no. October, pp. 1-15, 2020, doi: 10.48550/ARXIV.2010.09337.

[12] N. Rane, S. Choudhary, and J. Rane, "Explainable Artificial Intelligence (XAI) Approaches for Transparency and Accountability in Financial Decision-Making," *SSRN Journal*, vol. 2023, no. December, pp. 1-17, 2023, doi: 10.2139/ssrn.4640316.

[13] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable Machine Learning in Credit Risk Management," *Comput Econ*, vol. 57, no. 1, pp. 203–216, Jan. 2021, doi: 10.1007/s10614-020-10042-0.

[14] C. Onay and E. Öztürk, "A review of credit scoring research in the age of Big Data," *JFRC*, vol. 26, no. 3, pp. 382–405, July 2018, doi: 10.1108/JFRC-06-2017-0054.

[15] D. Martens, B. Baesens, T. Van Gestel, and J. Vanthienen, "Comprehensible credit scoring models using rule extraction from support vector machines," *European Journal of Operational Research*, vol. 183, no. 3, pp. 1466–1476, 2007, doi: 10.1016/j.ejor.2006.04.051.

[16] P. Adler et al., "Auditing black-box models for indirect influence," *Knowl Inf Syst*, vol. 54, no. 1, pp. 95–122, Jan. 2018, doi: 10.1007/s10115-017-1116-3.

[17] S. Barocas and A. D. Selbst, "Big Data's Disparate Impact," *SSRN Journal*, vol. 2014, no. August, pp. 1-62, 2016, doi: 10.2139/ssrn.2477899.

[18] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint*, vol. 2017, no. February, pp. 1-13, 2017, doi: 10.48550/arXiv.1702.08608.

[19] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. January, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.

[20] S. Tyagi, "Analyzing Machine Learning Models for Credit Scoring with Explainable AI and Optimizing Investment Decisions," vol. 5, no. 01, pp. 1-16, 2022, doi: 10.48550/ARXIV.2209.09362.

[21] J. Černevičienė and A. Kabašinskas, "Explainable artificial intelligence (XAI) in finance: a systematic literature review," *Artif Intell Rev*, vol. 57, no. 8, p. 216, July 2024, doi: 10.1007/s10462-024-10854-8.

[22] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in AI," *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, vol. 2019, no. January, pp. 279–288, 2019, doi: 10.1145/3287560.3287574.

[23] C. Wang and Z. Xiao, "A Deep Learning Approach for Credit Scoring Using Feature Embedded Transformer," *Applied Sciences*, vol. 12, no. 21, p. 10995, Oct. 2022, doi: 10.3390/app122110995.

[24] S. Kakkar, "EXPLAINABLE AI MODELS FOR CREDIT RISK SCORING IN BANKING: BALANCING ACCURACY AND REGULATORY TRANSPARENCY," *IJFDS*, vol. 3, no. 2, pp. 1–6, Aug. 2025, doi: 10.34218/IJFDS_03_02_001.

[25] Xiaoyang Meng, Ying Jin, "Enhancing Corporate Financial Transparency and Performance Assessment through Big Data and Machine Learning," *J. Comb. Math. Comb. Comput.*, vol. 127a, no. April, pp. 6893-6908, Apr. 2025, doi: 10.61091/jcmcc127a-383.